A Study on Techniques for Extracting Rare Itemsets

B.Kiruthika[1], S. Nithya Roopa[2]

[1]M.E – Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore-641049,

[2]Assistant Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore-641049,

Abstract: ***Itemset mining is an elementary data mining approach predominantly used for exploring interrelations among data. Frequent itemset mining exhibit interrelations frequently containing in data in which each of the items in the database consists of distinct weights. Mining infrequent itemsets are more intriguing and are used to attenuate cost function. Infrequent pattern mining is a challenging venture because there is a huge number of such patterns that can be derived from a given dataset. This paper concentrates on distinct techniques used for mining infrequent itemsets.***

Keywords: Data Mning, Association rules, Frequent itemsets, Rare itemsets

1. Introduction

A bulk amount of data is present in the Information Industry. It is fundamental to analyze the large amount of data and obtain useful information from it.

1.1 Data Mining

Data mining is the process of identifying interesting patterns and knowledge from huge amount of data. The process of discovering knowledge from data involves Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Mined data can be used in many applications like Production Control, Science Exploration.

1.2 Frequent Itemset Mining

The first attempt for itemset mining was directed on detecting frequent itemsets. An itemset is said to be frequent itemset if its number of occurrence in the inspected data is higher than the provided threshold.

Applications of Frequent itemset mining:
1. Biological data analysis
2. Medical Image Processing
3. Market Based Analysis

1.3 Infrequent Itemset Mining

An itemset is said to be infrequent itemset if its number of occurrence in the inspected data is equal to or lower than the maximum threshold. Applications of Infrequent itemset mining:
1. Bioinformatics
2. Fraud detection
3. Statistical disclosure of risk assessment from census data

1.4 Problems in mining infrequent itemsets are

1. How to identify interesting infrequent itemsets.
2. .How to effectively find them in large database.

1.5 Association Rule Mining

Association rule mining (ARM) is a method of identifying interesting rules from database. Association rule is defined in the form of $X \rightarrow Y$.If a transaction contains itemset X then it will likely contain itemset Y as well.

For example, an association rule Cricket bat→Cricket ball shows that if someone buys cricket bat, he/she will likely buy Cricket ball too. Association Rule Mining uses support and confidence to measure the strength of association rules.

$$\text{Support}(X{=}{>}Y) = \frac{\text{Number of transactions containing both X and Y}}{\text{Total number of transactions}}$$

$$\text{Confidence }(X{=}{>}Y) = \frac{\text{Number of transactions containing both X and Y}}{\text{Number of transactions containing X}}$$

Example : Transactional dataset

TID	List of Items
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Association rules:

{I1,I2} => I5 Confidence=2/4=50%
{I1,I5}=>I2 Confidence=2/2=100%
{I2,I5}=>I1 Confidence=2/2=100%
I1=>{I2,I5} Confidence=2/6=33%
I2=>{I1,I5} Confidence=2/7=29%
I5=>{I1,I2} Confidence=2/2=100%

2. Related Work

2.1 On Minimal Infrequent Itemset Mining

Minimal Infrequent Itemset algorithm is used for mining minimal infrequent itemsets. An itemset is said to be minimal infrequent itemset if it is lesser than or equal to maximum threshold and does not contain any infrequent subset in the transactional dataset. MINIT algorithm is based upon the SUDA2 algorithm which is developed for finding minimal unique itemsets.

Items ranking is determined by estimating the support of each of the items.A list of items in increasing order of support is found. Minimal infrequent itemsets are detected by considering each item in rank order.

MINIT algorithm is called recursively considering only those items with higher rank. Each candidate's minimal infrequent itemset is checked against the original dataset.

2.2 Apriori Based: Mining Infrequent and Non-Present Item Sets from Transactional Data Bases

In this method, only infrequent item sets and non-present item sets are identified. Minimum negative count is considered for detecting infrequent itemsets. No need to consider the minimum negative count to detect non present itemsets.

Apriori property is applied in finding infrequent and non-present itemsets. If there exists a infrequent itemset, then all its supersets are considered as infrequent. Super sets are considered in to the solution as infrequent k- item sets. Non-present item sets are found directly from infrequent 1-item sets.

Infrequent patterns and non-present patterns are found out within one data base scan. By finding the infrequent 1-item sets, all the remaining sets of infrequent item sets and non-present item sets can be determined.

2.3 Incremental updating algorithm for rare itemsets on weighted condition

Many databases in the real world are dynamic in nature. Incremental updating is an approach to effectively update the frequent and infrequent itemsets from the dynamic database. Incremental updating algorithm is essential for mining infrequent itemsets in dynamic databases.

The MIIWIU algorithm is to discover frequent and rare itemsets when a new dataset is added in the original dataset and value of minimum support is not modified.

In this algorithm,
1) Improved Apriori algorithm is used. Each item in the transactional dataset is associated with a weight.
2) Used to discover frequent and infrequent itemsets of the original database, newly added database and the combined database of both original and newly added database.

MIIWIU algorithm addresses three main issues of incremental updating problem,
1) Discovering frequent and infrequent itemsets from the new database when a new dataset is inserted into the original database for a given minimum support and minimum confidence.
2) Generating frequent and infrequent itemsets from the new database when a new dataset is deleted from the old database for a given minimum support and minimum confidence.
3) Identifying frequent and infrequent itemsets when its minimum support and minimum confidence are changing.

2.4 Minimally Infrequent Itemset Mining using Pattern-Growth Paradigm and Residual Trees

IFP min algorithm is based on the pattern-growth model.It is used to discover minimally infrequent itemsets. MLMS model uses different thresholds for discovering frequent itemsets for distinct lengths of the itemset.IFP min algorithm executes by processing minimally infrequent itemsets by separating the dataset into two parts. One part contains a particular item and the other part does not contain the item. Items are sorted in each transaction in ascending order of their support counts.

Variation of the Apriori algorithm is used for extracting minimally infrequent itemsets. Residual trees are used to find multiple level minimum support itemsets. The use of residual trees reduces the computation time. The inverse FP-tree is a compressed representation of the whole transactional database.

The IFP min algorithm recursively mines the minimally infrequent itemsets by dividing the IFP-tree into two sub-trees:
1. Projected tree
2. Residual tree.
The IFP min algorithm considers the projected tree and residual tree of only the infrequent item. IFP min must be used at higher thresholds and for larger datasets.

2.5 Infrequent Weighted Itemset Mining Using Frequent Pattern Growth

Discovers data items which occurs rarely in the given transactional dataset. Each data item is associated with a weight and the weight represents each data item's importance with each transaction. This approach uses two algorithms to mine infrequent weighted itemsets.
1. The Infrequent Weighted Itemset Miner Algorithm.
2. The Minimal Infrequent Weighted Itemset Miner Algorithm.

Initially, for the given transactional dataset, weighting function is calculated. And then infrequent weighted itemset support is determined. Infrequent itemsets for the original transactional dataset are generated based upon the support and threshold value. Equivalence weighted transactional dataset is generated for the original dataset and infrequent weighted itemsets for the equivalence dataset are identified .IWI Miner algorithm combines infrequent itemsets from both dataset and produces infrequent weighted itemsets. MIWI Miner algorithm is used to discover all minimally infrequent weighted itemsets in the given dataset.

2.6 Confabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets

CARM algorithm uses a cogency inspired measure for producing rules. Using this algorithm, rules are detected only by one file scan. Rule mining is performed in two steps:

1. Knowledge acquisition and structure construction.
2. Rule generation by confabulation and cogency measure.

Only one item consequent association rules are generated by this algorithm. Items are considered as symbols in this algorithm. After finding all frequent itemsets, algorithm generates all rules using their support and confidence. Knowledge acquisition consists of two modules. Axonal communication links between the two modules are used to store all domain knowledge. Rule extraction is completed based on the strength of the communication links.

2.7 Comparison

S	*Author*	*Algorithms*	*Merits*	*Demerits*
1	D. J. Haglin and A.M. Manning.	Minimal Infrequent Itemset algorithm.	Improved Performance is achieved.	Increased running time is observed.
2	Sujatha Kamepalli, Raja Sekhara Rao Kurra and Sundara Krishna.Y.K.	Apriori based algorithm.	Effectively mines infrequent and non-present itemsets within one database scan.	No pruning strategy is implemented and hence complexity must be improved.
3	Wenjuan Dong, He Jiang, Lei Chen, and Guoling Liu.	MIIWIU algorithm.	Scalable and efficient in identifying correlations among weighted Itemsets.	Not reliable for high dimensional data.
4	A. Gupta, A. Mittal, and A. Bhattacharya.	IFP min algorithm.	Higher Performance is achieved.Computational time is reduced	Scalability is not improved.
5	Luca Cagliero and Paolo Garza.	1.IWI Miner algorithm. 2.MIWI Miner algorithm.	Reduces the computational time.Performance is improved.	Algorithm must be improved to use in advanced decision making system.
6	Azadeh Soltani and M.-R. Akbarzadeh-T.	CARM algorithm.	Much faster than apriori algorithm.Improved classification error rate.	Efficiency of CARM's implementation must be improved.

3. Conclusion

The presented work surveys various methods for extracting infrequent item sets of data. The major advantage for discovering infrequent itemset was to advance the profit of rarely originated datasets in the transactions. Merits and demerits of each method are discussed to effectively differentiate the each methods functionality.

References

[1] D. J. Haglin and A.M. Manning, “On Minimal Infrequent ItemsetMining,” Proc. Int‟l Conf. Data Mining (DMIN ‟07), pp. 141-147, 2007.
[2] Sujatha Kamepalli, Raja Sekhara Rao Kurra and Sundara Krishna.Y.K, “Apriori Based: Mining Infrequent and Non-Present Item Sets from Transactional Data Bases,” International Journal of Electrical & Computer Science IJECS-IJENS Vol:14 **No:03.**
[3] Wenjuan Dong, He Jiang, Lei Chen, Guoling Liu, “Incremental updating algorithm for infrequent itemsets on weighted condition,” 2010 International Conference On Computer Design And Appliations.
[4] A. Gupta, A. Mittal, and A. Bhattacharya, “Minimally Infrequent Itemset Mining Using Pattern-Growth Paradigm and Residual Trees,” Proc. Int‟l Conf. Management of Data (COMAD), pp. 57-68, 2011.
[5] Luca Cagliero and Paolo Garza, “ Infrequent Weighted Item set Mining Using Frequent Pattern Growth” IEEE Transactions On Knowledge And Data Engineering Volume 26, No.4, April 2014.
[6] Azadeh Soltani and M.-R. Akbarzadeh-T., *Senior Member,IEEE,* “Confabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets”, IEEE transactions on neural networks and learning systems, vol.25,no.11,November2014.

Author Profile

B. Kiruthika, ME (Computer Science & Engineering) Kumaraguru College of Technology, India.

S. Nithya Roopa, Assistant Professor (Department of Computer Science & Engineering) Kumaraguru College of Technology, India.